Communicating generalizations about events

Michael Henry Tessler (mtessler@stanford.edu) and **Noah D. Goodman** (ngoodman@stanford.edu)
Department of Psychology, Stanford University

Abstract

Habitual sentences (e.g. *Bill smokes.*) generalize an event over time, but how do you know when a habitual sentence is true? We develop a computational model and use this to guide experiments into the truth conditions of habitual language. In Expts. 1 & 2, we measure participants' prior expectations about the frequency with which an event occurs and validate the predictions of the model for when a habitual sentence is acceptable. In Expt. 3, we show that habituals are sensitive to top-down moderators of expected frequency: It is the expectation of future tendency that matters for habitual language. This work provides the mathematical glue between our intuitive theories' of others and events and the language we use to talk about them. **Keywords:** events; generics; pragmatics; Bayesian data analysis; Bayesian cognitive model

Figuring out that a person or thing tends to exhibit a behavior can be crucial and hard-won knowledge. It is not surprising then that language has ways to convey this information about propensity: *Bill smokes*, *That man steals from children*, *My car doesn't start*. These are called *habitual sentences*. Like many other linguistic means of conveying generalizations, the truth-conditions of habituals are extremely flexible: If Bill smoked three cigarettes last month, can you say that *Bill smokes*? If he smoked a pack last week, but just swore a solemn oath to quit? In this paper we present the first empirical data on the felicity of habitual sentences and describe a formal model of habitual interpretation, based on recent Bayesian models of the pragmatics of vague language.

Linguists have pointed out the parallel between habituals like *Bill smokes* and generic sentences like *Swans are white* (Carlson, 1977, 2005; Cohen, 1999). Both convey generalizations (about events and categories, respectively), and both exhibit dramatic flexibility in their truth conditions: *Swans are white* even though there are black swans, and it may the case that *Bill smokes* even if he goes without a cigarette for an entire family vacation. Indeed, cases like *Mosquitos carry malaria* (wherein only a tiny percentage have the property) seem to parallel habitual sentences of rare actions like *Susan writes novels*. Susan may only have written 3 novels in her life, but still this seems like a valid generalization to convey.

In this paper, we take the analogy between generic and habitual language seriously by elaborating a recent computational theory of generic language to derive predictions for the truth conditions of habitual sentences. The theory of Tessler and Goodman (under review) posits that the semantics of a generic statement is an uncertain threshold on the degree of property prevalence (i.e. how many instances of the kind have the feature) and derives context-sensitive meanings through pragmatic inference given the distribution of property prevalence (i.e. in general, what prevalences are likely across different categories). We adapt this theory to explain habituals by adjusting the underlying degree to be the *propensity to take part in an event* (e.g. how often does a person tend to do an action) and derive predictions for felicity judgments of a range of habituals. In Expt. 1, we measure *a priori* beliefs about how frequently people do a diverse set of actions. In Expt. 2, we use those priors and the pragmatic theory to make predictions about the truth conditions of habitual sentences under different frequencies of action.

If habituals (and generics) are truly language for conveying generalizations, it would be useful for them to reflect expectations, not merely observations. Is the underlying dimension for habituals the actual, past frequency of the event (e.g. how often Bill has actually smoked in the past week) or the predictive probability that this event will occur in the near future (e.g. how likely it is that Bill will smoke next week)? In Expts. 3a & 3b, we show that the object of communication is a speaker's prediction about the future frequency of action, and not past frequency. This has important implications about the relationship between linguistic generalizations and intuitive theories, which we explore briefly in the discussion.

Computational model

A habitual sentence expresses a generalization about the tendency of an individual to participate in a kind of event. For a given individual (e.g. BILL) and an event or behavior (e.g. SMOKING), we refer to the rate at which the individual participates in the event for a given time window as the *propensity*, denoted λ. A natural denotation for the habitual is a simple threshold on propensity $\lambda > \tau$ (c.f. Cohen, 1999), yet no fixed value for τ would lead to the observed flexibility of truth conditions (e.g. *Bill smokes* vs. *writes novels*). Building on Lassiter and Goodman (2013, 2015), we posit that this threshold is not a fixed property of the language, but is established by pragmatic inference.

We model a speaker S_2 who reasons about a pragmatic listener L_1; this listener is considering the propensity of a certain behavior of an individual. The listener L_1 has uncertainty about the appropriate threshold for the habitual in this context ($\tau \sim$ Uniform over possible frequencies), and reasons about what an informative speaker S_1 would be likely to say. The hypothetical speaker S_1 in turn reasons about an idealized literal listener L_0, who has access to the threshold τ (i.e. S_1 believes L_0 will interpret him in exactly the way he means). Writing the propensity as λ, this leads to a set of equations:

$$P_{S_2}(u \mid \lambda) \propto \exp(\alpha_2 \cdot \ln \int_\tau P_{L_1}(\lambda, \tau \mid u)\, d\tau) \quad (1)$$

$$P_{L_1}(\lambda, \tau \mid u) \propto P_{S_1}(u \mid \lambda, \tau) \cdot P(\lambda) \cdot P(\tau) \quad (2)$$

$$P_{S_1}(u \mid \lambda, \tau) \propto \exp(\alpha_1 \cdot \ln P_{L_0}(\lambda \mid u, \tau)) \quad (3)$$

$$P_{L_0}(\lambda \mid u, \tau) \propto \delta_{[\![u]\!](\lambda,\tau)} P(\lambda). \quad (4)$$

We take the speakers S_1 and S_2 to consider two utterances: the habitual, with $[\![u]\!](\lambda,\tau) := \lambda > \tau$, or nothing (staying silent), with $[\![u]\!](\lambda,\tau) :=$ True, and to select utterances softmax optimally, with a degree of optimality governed by α_1 and α_2, respectively. Equation 1 can then be interpreted as a model of felicity or truth judgments (Degen & Goodman, 2014; Tessler & Goodman, under review). The speaker will choose to produce the habitual when the true propensity λ is more likely under L_1's posterior given the habitual than under her prior (implied by S_2 "staying silent"). A fully implemented version of the model can be found at `http://forestdb.org/models/habituals-cogsci2016.html`.

The prior, $P(\lambda)$ in Eqs. 2 and 4, specifies prior beliefs about the propensity of a specific event or behavior (e.g. SMOKING) across a set of different individuals. *A priori*, it is unclear how far to extend this *contrast set*: Does it include beliefs about *all* people, or just individuals within a predefined subclass e.g. individuals of the same gender or the same age? This is important because the meaning of the habitual (the threshold τ) is derived with respect to the prior. If the priors differed by gender (e.g., in the propensity to WEAR A BRA) and if language interpreters took the prior to be with respect to a particular gender, the model would predict differences in the truth conditions by gender (e.g., in *Susan wears a bra.* vs. *Bill wears a bra.*). We explore this issue in Expts. 1 & 2.

Experiment 1: Prior elicitation

In this experiment we elicit the prior $P(\lambda)$ for different events in order to generate model predictions for corresponding habituals. Given that some individuals rarely or never engage in an event, while others do quite frequently, we would expect the prior to be a mixture distribution between (at least) these two possibilities, similar in spirit to Zero-inflated or Hurdle Models of epidemiological data (Rose, Martin, Wannemuehler, & Plikaytis, 2006). Indeed, there may be more than these two possibilities, corresponding to individuals with different traits or demographics (e.g., different expected frequencies depending on age or gender).

Methods

Participants We recruited 40 participants from Amazon's Mechanical Turk. Participants were restricted to those with U.S. IP addresses and who had at least a 95% work approval rating. The experiment took on average 12 minutes and participants were compensated $1.25 for their work.

Materials We created thirty-one events organized into pairs or triplets from 5 different conceptual categories: food and drug (e.g. *eats caviar*, *eats peanut butter*), work (e.g. *sells things on eBay*, *sells companies*), clothing (e.g. *wears a suit*, *wears a bra*), entertainment (e.g. *watches professional football*, *watches space launches*) and hobbies (e.g. *runs*, *hikes*). Items were chosen to intuitively cover a range of likely frequencies of action, as well as to provide a minimal comparison to another item by having a common superordinate action (e.g. *eating* caviar vs. peanut butter).

Procedure For each event, participants were asked two questions, with associated dependent measures:

1. "How many {men, women} have DONE ACTION before?"
   Participants responded "N out of every J." by entering a number for N and choosing J from a drop-down menu (options: {1000 - 10 million}; default: 1000).
2. "For a typical {man, woman} who has DONE ACTION before, how frequently does he or she DO ACTION?"
   Participants responded "M times in K." by entering a number for M and choosing K from a drop-down menu (options: {week, month, year, 5 years}; default: year).

For example, one set read: "How many men have smoked cigarettes before?"; "For a typical man who has smoked cigarettes before, how frequently does he smoke cigarettes?"

We anticipated there might be different beliefs about the frequency of events depending on whether the actor is male or female, so we asked about both genders. Participants answered both questions for each gender on each slide (4 questions total per slide, order of male / female randomized between-subjects), and every participant completed all 31 items in random order. The experiment in full can be viewed at `http://stanford.edu/~mtessler/habituals/experiments/priors/priors-2.html`.

Data analysis and results

We built a Bayesian data analysis model for this prior elicitation task. Question 1 elicits the proportion of people who have done an action before. We model this data as coming from a Beta distribution: $d_1 \sim \text{Beta}(\gamma_1, \xi_1)$. Question 2 elicits the rate, or relative frequency, with which a person does the action. This was modeled by a log-normal distribution: $\ln d_2 \sim \text{Gaussian}(\mu_2, \sigma_2)$. Each item was modeled independently for each gender. We implemented this model using the probabilistic programming language WebPPL (Goodman & Stuhlmüller, 2014), and found the credible values of the parameters by running MCMC for 100,000 iterations, discarding the first 50,000 for burnin.

The priors elicited cover a range of possible parameter values as intended (Figure 1, scatter), resulting in parametrized distributions of dramatically different shapes (insets). We observe a correlation in our items between the mean % of Americans who have DONE ACTION before (Question 1) and the mean log-frequency of action (Question 2) ($r_{1,2} = 0.74$). Items that tend to be more popular actions also tend to be more frequent actions (e.g. *wears socks*) and visa-versa (e.g. *steals cars*), though there are notable exceptions (e.g. *plays the banjo* is not popular but done frequently when done at all, as is *smokes cigarettes*; *goes to the movies* is a popular activity though not done very often). This diversity is relevant because the speaker model (Eq. 1) will produce habitual sentences (e.g. *Sam goes to the movies vs. the ballet.*) contingent on the shape of the prior distribution.

From the inferred parameters and assumed functional forms, we get an inferred $P(\lambda)$ modeled as a mixture of individuals with the possibility of carrying out the action and those without the possibility of doing it. That is, $P(\lambda)$ was

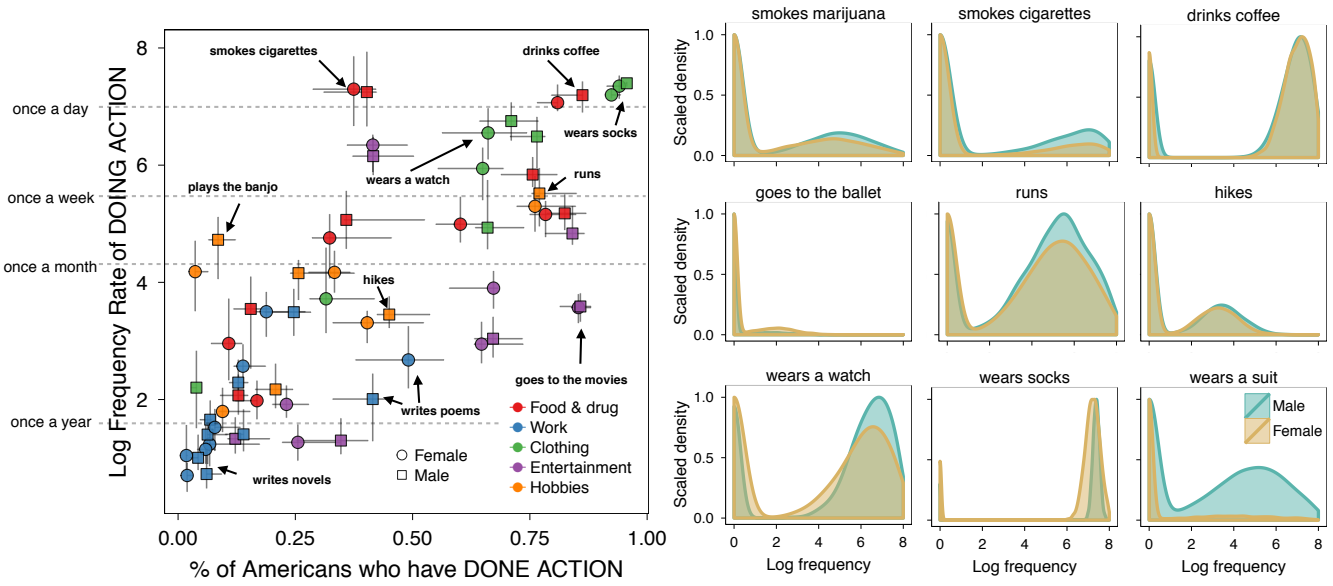

Figure 1: Frequency prior distributions empirically elicited for thirty-one events for both male and female genders. Left: Prior distributions are summarized by θ — the proportion of people who have done the action before — and γ — the mean log frequency of doing the action (for people who had done it before). Error bars denote 95% Bayesian credible intervals. Right: Density plots display posterior predictive distributions on frequency using the structured Bayesian model in Eq. 5. Log frequency scale is on the order of 5 years. Approximate rates of once per: {year $\sim$ 1.5; month $\sim$ 4; week $\sim$ 5.5; day $\sim$ 7}.

constructed by sampling λ as follows:

$$\theta \sim \text{Beta}(\gamma_1, \xi_1)$$
$$\ln\lambda \sim \begin{cases} \text{Gaussian}(\mu_2, \sigma_2) & \text{if Bernoulli}(\theta) = \text{T} \\ \delta_{\lambda=-\infty} & \text{if Bernoulli}(\theta) = \text{F} \end{cases} \quad (5)$$

In addition to specifying the correct way to combine our two prior-elicitation questions, using this inferred prior in our language model resolves two technical difficulties: (1) It smooths effects that are clearly results of the response format[1] and (2) it better captures the tails of the prior distribution which have relatively little data and need to be regularized by the analysis. Figure 1 (right) shows example inferred priors.

Some items show substantial differences between the genders (e.g. *wears a bra*) and some show subtle differences (e.g. *watches professional football*). We will explore the possibility of different truth conditions for habituals of different gendered characters in Experiment 2, for select items with priors that differ substantially by gender.

Experiment 2: Felicity judgments

A present-tense habitual sentence is of the form SINGULAR NOUN PHRASE + PRESENT TENSE SIMPLE VERB PHRASE (e.g. *Bill smokes cigarettes.*). We next explore the endorsements of habituals of this form made from the items whose propensity priors were measured in Experiment 1.

[1]For example, a very common rating is *1 time per year*. Presumably participants would be just as happy reporting *approximately* 1 time per year; the raw data does not reflect this due to demands of the dependent measure.

Methods

Participants We recruited 150 participants from MTurk. To arrive at this number, we performed a Bayesian precision analysis to determine the minimum sample size necessary to reliably ensure 95% posterior credible intervals no larger than 0.3 for a parameter whose true value is 0.5 and for which the data is a 2 alternative forced choice. This analysis revealed a minimum sample size of 50 per item; since participants only completed about one third of the items, we recruited 150 participants. The experiment took 4 minutes on average and participants were compensated $0.55 for their work.

Procedure and materials On each trial, participants were presented with a *past frequency statement* for a given event of the form: "In the past M {weeks, months, years}, PERSON DID X 3 times". For example, *In the past month, Bill smoked cigarettes 3 times*. The particular intervals used (number M and window {weeks, months, years}) were selected after examining the predictions of the speaker model (Eq. 1), for each item independently, to yield a variety of predicted endorsement rates. The items were the same as in Expt. 1.

Participants were asked whether they agreed or disagreed with the corresponding habitual sentence: "PERSON DOES X" (e.g. *Bill smokes cigarettes*). Participants saw 25 out of the 31 items paired randomly with a male or female character name; the other 6 trials were presented with both male and female names (on separate trials; 37 trials total) to explore the nature of the contrast class (see Model section). The experiment in full can be viewed at `http://stanford.edu/~mtessler/habituals/experiments/truth-judgments/tj-2.html`.

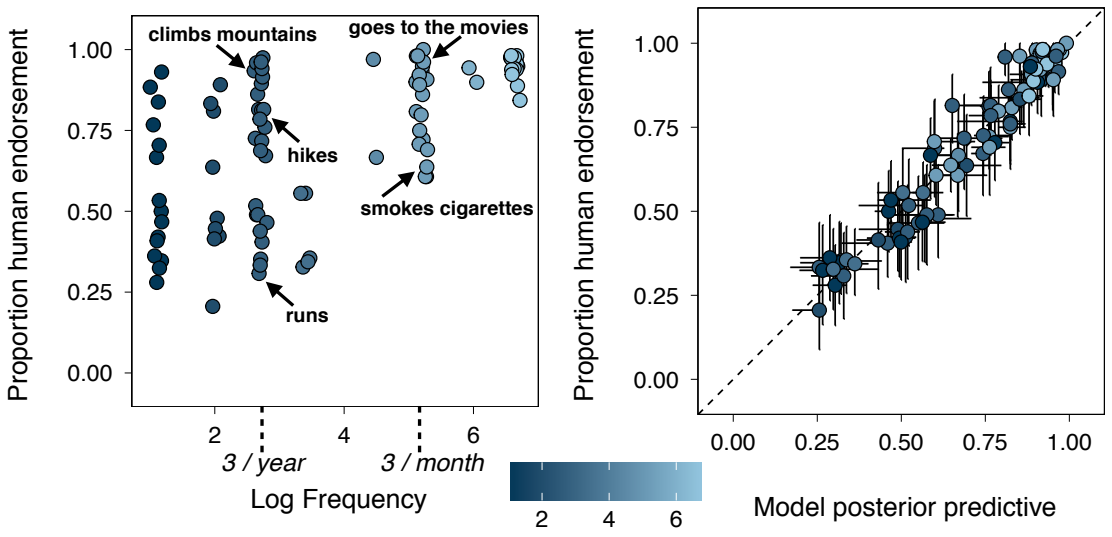

Figure 2: Human acceptability judgments as a function of the log frequency of action (left) and speaker S_2 model predictions (right) for ninety-three unique items (event x frequency). Color denotes target-individual frequency of action (log scale), with lighter colors indicating more frequent actions. Actual frequency noted on x-axis for examples (left). Error bars correspond to 95% bootstrapped confidence intervals for the participant data and 95% Bayesian credible intervals for the model predictions. Error bars suppressed and points jittered on left facet for visual clarity.

Results

Behavioral results On each trial of the experiment, the participant was told a person did a particular action 3 times during some time window. Figure 2 (left) shows the correspondence between the frequency of the event (normalizing to a 5-year time scale and taking the logarithm) and the felicity of the corresponding habitual sentence. It is clear that a habitual sentence can receive strong agreement even when the actions are very infrequent (log frequency $\sim$ 1; 3 times in a 5-year interval; e.g. *writes novels*, *steals cars*). We also see even when actions are done relatively frequently (e.g. 3 times in a one month interval; log frequency $\sim$ 5), there are habitual sentences participants are reluctant to endorse completely (e.g. *wears socks*, *drinks coffee*). In our data, actions completed with a high frequency (3 times in a one week interval; log frequency $\sim$ 6.5) receive at least 75% endorsement, though there is still variability among them (e.g. between 10-25% of people disagree with *wears a watch* and *wears a bra*). Overall, frequency of action predicts only a fraction of the variability in responses ($r^2(93) = 0.33$). For actions that are done on the time scale of years or longer (lower median of frequency), frequency itself no longer explains the endorsements ($r^2(50) = 0.07$)

We further examined the six items for which we observed gender differences in the prior elicitation task (Expt. 1). We find no differences between endorsements of the habitual of characters with male and female names, and overall, the mean endorsements by gender are strongly correlated $r(93) = 0.91$. This may be because the felicity of habitual sentences depends on a comparison to individuals of both genders (i.e, the *contrast class* is other people; not just other men or women). Less interestingly, the lack of a difference may be the result of gender being not very salient in our paradigm, perhaps because the names used were not sufficiently gendered.

Model fit and results We used the pragmatic speaker model S_2 (Eq. 1) with the priors elicited above (Expt. 1) to predict felicity judgments in Expt. 2, assuming the target propensity (to be conveyed by S_2) is the provided frequency. Because we observe no difference between the felicity judgments for habituals of male and female characters, we use a 50% mixture of the inferred priors for each gender to construct a single frequency distribution $P(\lambda)$ across individuals. The model has two free parameters—the speaker optimality parameters, α_i, in Eqs. 1 & 3. We use Bayesian data analytic techniques to integrate over these parameters (Lee & Wagenmakers, 2014), comparing the posterior predictive distribution to the empirical data in Expt. 2. To construct the posterior predictive distribution over responses, we collected 2 MCMC chains of 100,000 iterations, discarding the first 50,000 iterations for burn in. The Maximum A-Posteriori (MAP) value and 95% highest probability density interval (HDI) for α_1 is 19.3 [14.9,19.9] and α_2 is 1.5 [1.4,1.6].

As shown in Figure 2, right, the probabilistic pragmatics model does a good job of accounting for the variability in responses ($r^2(93) = 0.94$), including actions done on the time scale of years or more ($r^2(50) = 0.92$). The model decides when the habitual is a useful way to describe the person's behavior, assuming that what the person did in the past is representative. This raises an interesting question: Does the propensity communicated by the habitual indicate an objective, past frequency or a subjective, future expectation?

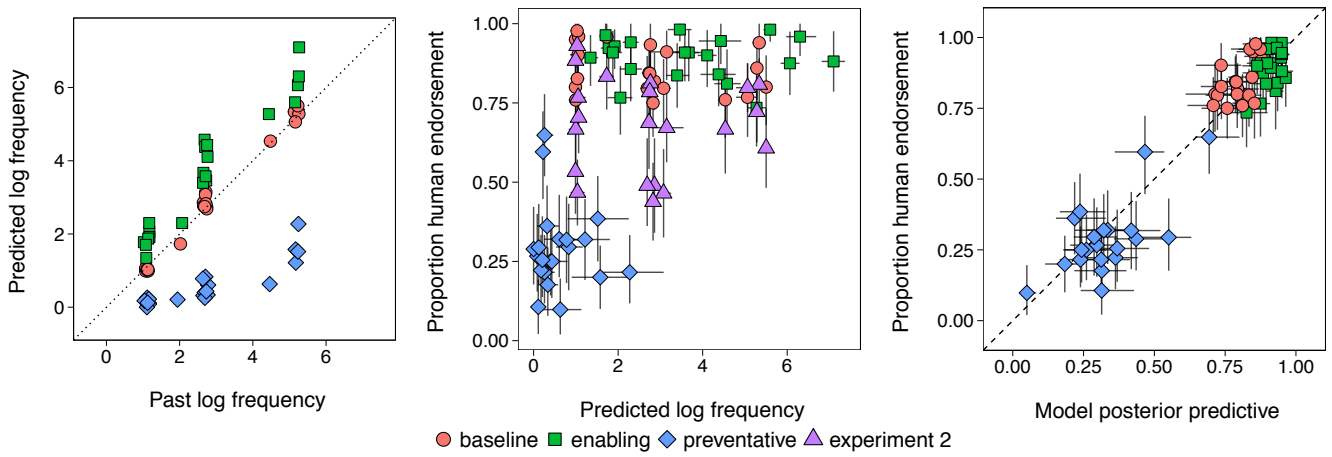

Figure 3: Left: Predicted log frequency as a function of past log frequency given to the participant (Expt. 3a; CIs suppressed and jitter added for visual clarity). Middle: Human endorsements of habitual sentences (Expt. 3b) vs. Predicted log frequency (Expt. 3a), with data for corresponding items from Expt. 2 (assumed to have the same predictive log frequency as baseline). Right: Endorsements (Expt. 3b) vs. Speaker S_2 model predictions using empirically elicited predictive frequencies (Expt. 3a).

Experiment 3: Objective frequency versus subjective expectation

While past frequency is often a good indicator of future tendency, people can change abruptly due to a variety of decisions and outside events. Does habitual language communicate propensity in terms of past frequency or future expectations? In this set of experiments, we address this by introducing causal factors that enable or prevent future actions (e.g. buying cigarettes; developing an allergy). In Expt. 3a, we measure *predictive frequency* when past frequency alone is observed and when these causal factors are introduced. In Expt. 3b, we examine felicity judgments of the habitual sentence (e.g. *John smokes cigarettes.*; *John eats peanut butter.*) in the presence of these causal modifiers. This will allow us to test whether habituals are best explained by a speaker S_2 who communicates the known past or expected future frequency.

Experiment 3a: Prediction elicitation

Methods We recruited 120 participants from MTurk, using the same criterion as Expt. 2. The experiment took 4 minutes on average and participants were compensated $0.40.

The procedure was identical to Expt. 2 except for the inclusion of a second sentence on a subset of trials and the use of a different dependent measure. On all trials, participants were presented with a *past frequency sentence* (see Expt. 2). Additionally, on one third of the trials, participants were presented with a **preventative sentence** (e.g. *Yesterday, Bill quit smoking.*). On one third of the trials, participants were presented with an **enabling sentence** (*Yesterday, Bill bought a pack of cigarettes.*) The final third of trials had no additional evidence and were identical to Expt. 2.

Only twenty-one of the original thirty-one items were used in order to shorten the experiment. To increase expected variability, participants saw only the frequencies that led to most intermediate endorsement of the habitual in Expt. 2. In addition, we did not include separate trials for both male and female names for the select items we did in Expt. 2, since we saw no differences in their endorsements of the habitual.

Participants were asked "In the next TIME WINDOW, how many times do you think PERSON does EVENT?", where the TIME WINDOW was the same as given in the *past frequency statement*. The experiment in full can be viewed at `http://stanford.edu/~mtessler/habituals/experiments/priors/predictive-1.html`.

Behavioral results

Figure 3 (left) shows the predicted future frequency as a function of the past frequency given to the participant and the type of causal information given. We observe in the baseline condition that future frequency perfectly tracks past frequency (e.g. participants believe if a person smoked cigarettes 3 times last month, they will smoke cigarettes 3 times next month). This means that our model makes identical predictions for Expt. 2 whether the target is past frequency or expected future frequency (indicating, as expected, that we must look to the new data to distinguish these models). Critically, we observe the preventative information appreciably decreasing and the enabling information slightly increasing predicted frequency (Figure 3 left; blue and green dots).

Experiment 3b: Felicity judgments

Methods We recruited 150 participants from MTurk, using the same criterion as Expt. 2. The experiment took 4 minutes on average on participants were compensated $0.40 for their work. None of the participants had completed Expt. 3a.

The only difference from Expt. 3a is the dependent measure. On each trial, participants were asked if they agreed or disagreed with the corresponding habitual sentence (as in Expt. 2).

The experiment in full can be viewed at `http://stanford.edu/~mtessler/habituals/experiments/truth-judgments/tj-3-preventative.html`.

Results

There is a clear and consistent negative effect of preventative information on endorsements for the habitual sentence (Figure 3, middle; blue points). When collapsing across items and subjecting the data to a generalized mixed-effects model with random by-participant effects of intercept and random by-item effects of intercept and conditions, we find evidence for a small effect of *enabling* conditions on endorsements (M = 0.89; 95% Bootstrapped CI [0.88, 0.91]) as compared to baseline (M = 0.85 [0.83, 0.87]) [$\beta = 0.42; SE = 0.15; z = 2.8$], and a large effect of *preventative* conditions on endorsements (M = 0.29 [0.26, 0.31]) [$\beta = -3.22; SE = 0.21; z = -15.2$].

We use the mean predicted log frequency from Expt. 3a as the input to the speaker S_2 model to predict the felicity judgments measured in Expt. 3b. We infer the one model parameter using the same analysis approach in Expt. 2. The model matches the data well ($r^2(63) = 0.91$; Figure 3, right). The same model using the past frequency as the object of communication does not match the data at all ($r^2(63) = 0.02$). These results suggest that the felicity of habituals is based on an underlying scale of predicted future propensity, not merely the observed frequency in the past.

Interestingly, we observe endorsements in this experiment that are appreciably higher than in Expt. 2 for the same items (Figure 3, middle; red vs. purple points). This may be due, in part, to an effect of the experimental context on participants: in this experiment the overall population of frequencies is much lower (both because we selected moderate frequencies from Expt. 2 and because of the preventative information) and participants may infer that the experimenter believes this to be a representative range and adjust judgments accordingly. Future investigation into this issue is warranted.

Discussion

We presented a computational model for communicating generalizations about events. The model decides if a habitual sentence is a pragmatically useful way to describe a person's behavior, taking into account the listener's prior beliefs about the action—how common it is and the likely frequency (measured in Expt. 1). We validated this model by eliciting felicity judgments for habitual sentences covering diverse activities with a wide range of experimentally manipulated frequencies of action (Expt. 2). We further investigated the nature of the underlying "propensity" scale by introducing enabling and disabling evidence, measuring the predicted future frequency (Expt. 3a) and using that, with the model, to predict the felicity of habitual sentences (Expt. 3b). To our knowledge, the experiments presented here are first empirical investigations into the truth conditions of habitual sentences and the first test of a formal model of habitual language.

The model we present here is almost identical to a model we have used to describe generic language (Tessler & Goodman, under review). The only difference is in the underlying scale: for generics, it is the *prevalence* of the property; for habituals, the *propensity* of the action. This provides a formal bridge between generalizations about categories (i.e. *generics*) and generalizations about events (i.e. *habituals*), a connection often noted in the linguistics literature (Carlson, 1977, 2005; Cohen, 1999). Generics often use a bare plural (e.g. *Bears like to eat ants.*) and don't lay claim to any well-defined set of individuals (many bears may not like to eat ants). Habituals use the simple present tense (e.g. *John smokes*) without any well-defined period of time (John may go many days without smoking). In both cases, a pragmatically inferred threshold on prevalence or propensity, respectively, explains the varying truth conditions of these kinds of sentences. Scales, and scalar representations, provide a simple and general quantitative way to express truth conditions.

Accurately predicting the environment is critical for survival and development. Habituals convey generalizations about events and are helpful for future predictions about events. For example, knowing that an event generally happens may be a useful abstraction for causal inference (e.g. Griffiths & Tenenbaum, 2005). Generalizations about people's behavior in particular—as we've investigated in this article—are important to understand, as they likely facilitate trait induction and essentialist beliefs (Gelman & Heyman, 1999). The computational model presented here provides a mathematical bridge between the way we talk about people's behavior and our intuitive theories of others and events.

References

Carlson, G. N. (1977). *Reference to kinds in english*. Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Carlson, G. N. (2005). Generics, Habituals and Iteratives. In *The encyclopedia of language and linguistics* (2nd ed.).

Cohen, A. (1999). Generics, Frequency Adverbs, and Probability. *Linguistics and Philosophy*, *22*.

Degen, J., & Goodman, N. D. (2014). Lost your marbles? the puzzle of dependent measures in experimental pragmatics. In *Proceedings of the 36th annual conference of the Cognitive Science Society.*

Gelman, S. A., & Heyman, G. D. (1999). Carrot-Eaters and Creature-Believers: The Effects of Lexicalization on Children's Inferences About Social Categories. *Psychological Science*(6).

Goodman, N. D., & Stuhlmüller, A. (2014). *The Design and Implementation of Probabilistic Programming Languages.* `http://dippl.org`. (Accessed: 2015-12-30)

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive psychology*, *51*(4), 334–84.

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Semantics and Linguistic Theory (SALT) 23* (pp. 587–610).

Lassiter, D., & Goodman, N. D. (2015). Adjectival vagueness in a bayesian model of interpretation. *Synthese*.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge Univ. Press.

Rose, C. E., Martin, S. W., Wannemuehler, K. A., & Plikaytis, B. D. (2006). On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *J. Biopharmaceutical Statistics*, *16*(4), 463–481.

Tessler, M. H., & Goodman, N. D. (under review). *A pragmatic theory of generic language.*